An Automatic Finite-Sample Robustness Metric: Can Dropping a Little Data Make a Big Difference?

Ryan Giordano (rgiordan@mit.edu)[1]
January 2022

[1]With coauthors Rachael Meager (LSE) and Tamara Broderick (MIT)

Dropping data: Motivation

All the time in data science, we:

- Gather + clean exchangeable data,
- Specify and fit a model, and
- Drawn a qualitative conclusion from your fit (e.g., based on the sign / significance of some estimated parameter).

Decisions are important: We want **trustworthy** conclusions.
Data / models not always perfect: We want **robust** conclusions.

Running example: ?, a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points.

Would you be concerned if you could **reverse your conclusion** by removing a **small proportion** (say, 0.1%) of your data?

Dropping data: Mexico Microcredit

Consider **?**, a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points.
The variable “Beta” estimates the effect of microcredit in US dollars.

	Beta (SE)
Original result	-4.55 (5.88)

Dropping data: Mexico Microcredit

Consider **?**, a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points.
The variable “Beta” estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original result	0	-4.55 (5.88)
Change sign	1	0.4 (3.19)

Dropping data: Mexico Microcredit

Consider **?**, a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points.
The variable "Beta" estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original result	0	-4.55 (5.88)
Change sign	1	0.4 (3.19)
Change significance	14	-10.96 (5.57)

Dropping data: Mexico Microcredit

Consider **?**, a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points.
The variable "Beta" estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original result	0	-4.55 (5.88)
Change sign	1	0.4 (3.19)
Change significance	14	-10.96 (5.57)
Change both	15	7.03 (2.55)

Dropping data: Mexico Microcredit

Consider **?**, a randomized controlled trial study of the efficacy of microcredit in Mexico based on 16,560 data points.
The variable "Beta" estimates the effect of microcredit in US dollars.

	Left out points	Beta (SE)
Original result	0	-4.55 (5.88)
Change sign	1	0.4 (3.19)
Change significance	14	-10.96 (5.57)
Change both	15	7.03 (2.55)

By removing very few data points ($15/16560 \approx 0.1\%$), we can reverse the qualitative conclusions of the original study!

Dropping data: Motivation

Would you be concerned if you could **reverse your conclusion** by removing a **small proportion** (say, 0.1%) of your data?

Dropping data: Motivation

Would you be concerned if you could **reverse your conclusion** by removing a **small proportion** (say, 0.1%) of your data?

Not always! But sometimes, surely yes.

Thinking without random noise can be helpful.

Suppose you have a farm, and want to know whether your average yield is greater than 170 bushels per acre. At harvest, you measure 200 bushels per acre.

- Scenario one: If your yield is greater than 170 bushels per acre, you make a profit.
  - Don't care about sensitivity to small subsets
- Scenario two: You want to recommend your farming methods to a friend across the valley.
  - Might care about sensitivity to small subsets

For example, often in economics:

- Small fractions of data are missing not-at-random,
- Policy population is different from analyzed population,
- We report a convenient summary (e.g. mean) of a complex effect,
- Models are stylized proxies of reality.

Objective

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Objective

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Question 1: How do we find influential datapoints?
The number of subsets $\binom{N}{\lfloor \alpha N \rfloor}$ can be very large even when α is very small.
In the MX microcredit study, $\binom{16560}{15} \approx 1.4 \cdot 10^{51}$ sets to check for $\alpha = 0.0009$.
We provide a fast, automatic approximation based on the **influence function**.

Objective

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Question 1: How do we find influential datapoints?

Question 2: What makes an estimator non-robust?

Objective

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Question 1: How do we find influential datapoints?

Question 2: What makes an estimator non-robust?
Non-robustness to removal of $\lfloor \alpha N \rfloor$ points is:

- Not (necessarily) caused by misspecification.
- Not (necessarily) caused by outliers.
- Not captured by standard errors.
- Not mitigated by large N.
- Primarily determined by the **signal to noise** ratio

  ... in a sense which we will define.

Objective

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Question 1: How do we find influential datapoints?

Question 2: What makes an estimator non-robust?

Question 3: When is our approximation accurate?

Objective

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Question 1: How do we find influential datapoints?

Question 2: What makes an estimator non-robust?

Question 3: When is our approximation accurate?

- We provide deterministic error bounds for small α.
- We show the accuracy in simple experiments.
- We show the accuracy in a number of real-world experiments.

Objective

Estimate the effect of leaving out $\lfloor \alpha N \rfloor$ datapoints, where α is small.

Question 1: How do we find influential datapoints?

Question 2: What makes an estimator non-robust?

Question 3: When is our approximation accurate?

Conclusion: Related work and future directions

Question 1:
How do we find influential datapoints?

Question 2:
What makes an estimator non-robust?

Question 3:
When is our approximation accurate?

Conclusion:
Related work and future directions